Outline

• Recap
• Tumor Phylogeny Inference from Bulk DNA-seq with Copy-Number Aberrations

Reading:

Infinite Sites Model = Two-state Perfect Phylogeny

The genome is large

Mutations are rare

Infinite sites model: multiple mutations never occur at the same position

Mutated Loci

<table>
<thead>
<tr>
<th>Species (cancer cells)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

1: mutated
0: not

All sites are bi-allelic: mutated or not.

[Kimura, 1969]
Infinite Alleles Model = Multi-state Perfect Phylogeny

**Infinite alleles model:**
- For any mutation, there are an infinite number of possibilities of what mutation looks like (states).
- So, the same position can be mutated multiple times, but it never mutates to the same “allele” or state.

**Site History:**
- Characters have integer states

Mutation Site
Infinite alleles model:
• For any mutation, there are an infinite number of possibilities of what mutation looks like (states).
• So, the same position can be mutated multiple times, but it never mutates to the same “allele” or state.

Characters have integer states
Infinite alleles model:
• For any mutation, there are an infinite number of possibilities of what mutation looks like (states).
• So, the same position can be mutated multiple times, but it never mutates to the same “allele” or state.

Site History:
Characters have integer states
Outline

• Recap

• Tumor Phylogeny Inference from Bulk DNA-seq with Copy-Number Aberrations

Reading:

Tumor Evolution as a Phylogenetic Tree

**Clonal Theory** [Nowell, 1976]
Observations are Leaves of a Perfect Phylogeny $T$

**Assumptions:**
- Mutations are single nucleotide variants (SNVs)
- Infinite sites assumption

Single-cell sequencing

$$M = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Binary Matrix $B$

$O(mn)$ [Gusfield, 1991]

Two-State Perfect Phylogeny Tree $T$

<table>
<thead>
<tr>
<th>Seq. method</th>
<th>Inferring $T$</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-cell</td>
<td>unmixed two-state perfect phylogeny</td>
<td>$O(mn)$</td>
</tr>
</tbody>
</table>
Observations are Mixtures of the Leaves of $T$

**Variant Allele Frequencies Matrix $F$**

$F = \begin{bmatrix}
0.4 & 0.0 & 0.0 & 0.0 & 0.3 & 0.2 \\
0.3 & 0.3 & 0.0 & 0.3 & 0.0 & 0.0 \\
0.4 & 0.4 & 0.4 & 0.0 & 0.0 & 0.0
\end{bmatrix}$

**Variant Allele Frequencies Matrix $F$**

**NP-complete**

AnceTree
[El-Kebir, Oesper et al. 2015]

Two-State Perfect Phylogeny Tree $T$

Mixing Proportions $U$

<table>
<thead>
<tr>
<th>Seq. method</th>
<th>Inferring $T$</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-cell</td>
<td><strong>unmixed</strong> two-state perfect phylogeny</td>
<td>$O(mn)$</td>
</tr>
<tr>
<td>bulk</td>
<td><strong>mixed</strong> two-state perfect phylogeny</td>
<td>NP-complete</td>
</tr>
</tbody>
</table>

TrAp [Strino et al., 2013], PhyloSub [Jiao et al., 2014]
CITUP [Malikic et al., 2015], BitPhylogeny [Yuan et al., 2015]
LICHeE [Popic et al., 2015], ...
Copy-Number Aberrations Confound VAFs

\[ F = \begin{bmatrix}
0.4 & 0.0 & 0.0 & 0.0 & 0.3 & 0.2 \\
0.3 & 0.4 & 0.0 & 0.3 & 0.0 & 0.4 \\
0.4 & 0.4 & 0.4 & 0.0 & 0.0 & 0.4 \\
\end{bmatrix} \]

SNVs

samples

0 : non-mutated
1 : mutated
2 : single-copy amplification
3 : ...

Need > 2 states:
Outline

• Multi-State Perfect Phylogeny Mixture Problem

• Combinatorial Characterization of Solutions

• Application to Cancer Bulk-Sequencing Data

• Results
Infinite Sites Generalizes to Infinite Alleles

**Two-State Perfect Phylogeny** – Infinite sites assumption: a character changes state once

**Frequency Matrix** $F$

$$
\begin{pmatrix}
S_1 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\
S_2 & 0.6 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\
S_3 & 0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4
\end{pmatrix}
$$

**Two-State PP Matrix** $B$

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 1
\end{pmatrix}
$$

**Two-State PP Tree** $T$
Infinite Sites Generalizes to Infinite Alleles

**Two-State Perfect Phylogeny** – Infinite sites assumption: a character changes state once

**VAF Factorization Problem (VAFFP):** Given $F$, find $U$ and $B$ such that $F = UB$ [El-Kebir, Oesper et al., 2015]
Infinite Sites Generalizes to Infinite Alleles

**Two-State Perfect Phylogeny** – *Infinite sites assumption*: a character changes state once

\[
F = U B
\]

**Multi-State Perfect Phylogeny** – *Infinite alleles assumption*: a character changes to a state once

\[
F = U B
\]

VAF Factorization Problem (VAFFP): Given \( F \), find \( U \) and \( B \) such that \( F = U B \) [El-Kebir, Oesper et al., 2015]

Multi-State PP Matrix \( A \)

Multi-State Perfect Phylogeny Tree \( T \)
Two-State Perfect Phylogeny

\[ F = U B \]

[El-Kebir, Oesper et al., 2015]

Multi-State Perfect Phylogeny

\[ F_i = U A_i \]

Multi-state Perfect Phylogeny Mixture Problem: Given \( F \), find \( U \) and \( A \) such that \( F_i = U A_i \) for all states \( i \)
Combinatorial Characterization of Solutions

**Two-State Perfect Phylogeny**

- **Frequency Matrix** $F$
  - $m$ samples
  - $n$ mutations
  - \[
  \begin{pmatrix}
  0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\
  0.6 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\
  0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4 
  \end{pmatrix}
  \]

- **Ancestry Graph** $G$
  - directed acyclic graph
  - mutations / characters
  - potential parental relationship

**Multi-State Perfect Phylogeny**

- **Frequency Tensor** $\mathcal{F}$
  - $p$
  - $q$
  - $c$
  - $d$
  - $k$ states
  - $n$ characters

**Theorem** [El-Kebir, Oesper et al., 2015; Popic et al., 2015]

Solutions are **spanning trees** that satisfy

\[
    f_{p,(c,1)} \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)} \quad (SC)
\]

**Theorem** [El-Kebir, Oesper et al., 2015]

VAFFP is NP-complete for $m = O(n)$
Combinatorial Characterization of Solutions

**Two-State Perfect Phylogeny**

<table>
<thead>
<tr>
<th>Sample</th>
<th>0</th>
<th>0</th>
<th>0.8</th>
<th>0.8</th>
<th>0.8</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Frequency Matrix $F$

![Directed acyclic graph for mutations/characters]

**Multi-State Perfect Phylogeny**

<table>
<thead>
<tr>
<th>Sample</th>
<th>0</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.6</th>
<th>0.7</th>
<th>0.0</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Frequency Tensor $F$

![Directed multi-graph for character-state pairs]

**Theorem** [El-Kebir, Oesper et al., 2015; Popic et al., 2015]

Solutions are **spanning trees** that satisfy

$$f_{p,(c,1)}(d,j) \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)}$$

(SC)

**Theorem** [El-Kebir, Oesper et al., 2015]

VAFFP is NP-complete for $m = O(n)$

**Theorem** [El-Kebir et al., 2016]

Solutions are **threaded spanning trees** satisfying

$$f_{p}^{+}(D_{(c,i)}) \geq \sum_{(d,j) \in \delta(c,i)} f_{p}^{+}(D_{(d,j)})$$

(MSSC)

**Theorem** [El-Kebir et al., 2016]

PPMDP is NP-complete even for $m = 2$ and $k = 2$
Outline

• Multi-State Perfect Phylogeny Mixture Deconvolution Problem

• Combinatorial Characterization of Solutions

• Application to Cancer Bulk-Sequencing Data

• Results
Application to Cancer Bulk Sequencing

Frequency Tensor $\mathcal{F}$

State Graph

# maternal copies

# mutated copies

# paternal copies

$\mathbf{F} = (\mathbf{x}, \mathbf{y}, \mathbf{z})$
Application to Cancer Bulk Sequencing

Frequency Tensor $\mathcal{F}$

- $m$ samples
- $c$ characters
- $k$ states

$\begin{align*}
{p} & : \text{# maternal copies} \\
{q} & : \text{# mutated copies} \\
{d} & : \text{# paternal copies}
\end{align*}$

$(x, y, z) = \begin{cases} 
(1, 1, 0) \\
(1, 1, 1) \\
(2, 1, 2)
\end{cases}$

Copy-number calls: THetA, Battenberg, ...

State Graph

- $x$: # maternal copies
- $y$: # paternal copies
- $z$: # mutated copies

Copy-number calls: THetA, Battenberg, ...

$\begin{align*}
\mu_{(1,1)} = 0.8 \\
\mu_{(2,1)} = 0.2 \\
h = 3/10 = 0.3
\end{align*}$

State Trees

- $S_1$: $0.54 \to 0.26 \to 0.2$
- $S_2$: $0.14 \to 0.66 \to 0.2$
- $S_3$: $0.34 \to 0.46 \to 0.2$
- $S_4$: $0.8 \to -0.46 \to 0.66$
### SPRUCE Enumerates Phylogenies

**Somatic Phylogeny Reconstruction Using Combinatorial Enumeration**

Available at: [http://compbio.cs.brown.edu/projects/spruce/](http://compbio.cs.brown.edu/projects/spruce/)

---

**Copy-number calls and mixing proportions**

- **(1, 1)**
  - \( \mu_{(1,1)} = 0.8 \)
- **(2, 1)**
  - \( \mu_{(2,1)} = 0.2 \)

**VAF**

\[ h = \frac{3}{10} = 0.3 \]

**VAF confidence interval**

\[ [h, \bar{h}] = [0.18, 0.32] \]

Error-Model for Variant Allele Frequencies
SPRUCE: Somatic Phylogeny Reconstruction Using Combinatorial Enumeration
Outline

• Multi-State Perfect Phylogeny Mixture Deconvolution Problem

• Combinatorial Characterization of Solutions

• Application to Cancer Sequencing

• Results
SPRUCE Accurately Recovers Simulated Trees

Parameters:
• # characters $n$: 5, 15
• # samples $m$: 2, 5, 10

Methods:
• SPRUCE
• PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity
SPRUCE Accurately Recovers Simulated Trees

Parameters:
• # characters $n$: 5, 15
• # samples $m$: 2, 5, 10

Methods:
• SPRUCE
• PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity
SPRUCE Accurately Recovers Simulated Trees

Parameters:
• # characters $n$: 5, 15
• # samples $m$: 2, 5, 10

Methods:
• SPRUCE
• PhyloWGS [Deshwar et al., 2015]

Increasing number of samples decreases ambiguity
More Samples vs. More Coverage
Violations of Infinite Alleles Assumption
Cancer Cell Fractions (CCFs) Cannot Be Inferred A Priori

CCFs are used extensively in studying intra-tumor heterogeneity and tumor evolution:

1. Timing of driver mutations

2. Tumor evolution and phylogeny reconstruction

3. Developmental patterns of metastases
Metastatic Evolution in Prostate Tumor

**Input:**
- 10 samples: whole-genome & targeted sequencing
- ~110 SNVs

**Tree building:**
1. Infer cancer cell fraction (CCF) for each SNV in each sample
2. Cluster SNVs by CCFs across samples
3. Construct tree using Pigeon-Hole-Principle (Sum Condition)


The evolutionary history of lethal metastatic prostate cancer
Cancer Cell Fractions Cannot Be Inferred A Priori

Compatible state trees and state frequencies

PPFIA1

S1

S2

Gundem et al.

Number of samples with distinct CCF intervals

42

0

1

2

3

4

TAS2R4
ENSG0000196960
GRIN2B
CCDC21
SENP6

0

⊿

0

⊿

2

⊿

4

⊿

6

⊿

8

1

⊿

2

⊿

4

⊿

PPFIA1

Cancer Cell Fraction

0.0

0.2

0.4

0.6

0.8

1.0

1.2

1.4

0

1

2

3

4

Number of samples with distinct CCF intervals

42

7

4

1

1

TAS2R4
ENSG0000196960
GRIN2B
CCDC21
SENP6

PPFIA1

Cancer Cell Fraction

S1

S2

Gundem et al.

0

⊿

0

⊿

2

⊿

4

⊿

6

⊿

8

1

⊿

2

⊿

4

⊿

0.27, 0.59

0.13, 0.45

0.28, 0.28

0.28, 0.28

[0.27, 0.59]

[0.13, 0.45]

[0.28, 0.28]

[0.28, 0.28]

[0.55, 0.72]

[0.36, 0.58]

[0.53, 0.95]
Conclusions

• Copy-number aberrations confound variant allele frequencies
  • SNVs and CNAs must be considered jointly in phylogeny reconstruction

• Generalization of infinite sites for SNVs is infinite alleles for SNVs + CNAs
  • Multi-state Perfect Phylogeny Mixture Problem (PPM)

• Complete combinatorial characterization of the problem
  • Solutions are constrained spanning trees in a directed multi-graph
  • PPM is NP-complete for $k = 2$ and $m = 2$

• Using combinatorial structure, SPRUCE accurately recovers simulated trees

• Cancer cell fractions cannot be uniquely inferred \textit{a priori} by considering SNVs in isolation

• Precise mathematical models are needed to describe evolutionary process in cancer